Statistical report

This document sets out our statistical model's assumptions, explains its implementation and presents the results.

Background

The aim of this analysis is to describe timecourse data about the density of CHO cell cultures that were given treatments that induce cell death through a process called apoptosis. Some cultures were given genetic interventions that aim to make them apoptosis-resistant, either by reducing the rate at which they die or by extending the period before they start to die. We would like to know which interventions have the most effect, and in what way.

Methods

Assumptions about the target system

We assume that in this scenario the cells exist in four states:

- R replicative, growing at a rate of $\mu R(t)$, where t is the current time and $R(t)$ is the current density of replicative cells
- Q_a growth arrest, transferring from normal at a rate of $k_q R(t)$
- Q_c death committed, transferring from growth arrest at a rate of $k_q R(t-\tau)$, where $tau > 0$ represents the delay between growth arrest and death commitment.
- dead, transferring from death committed at a rate of $k_d Q_c(t)$, where $Q_c(t)$ is the density of death-committed cells at time t.

These assumptions define a system of ordinary differential equations (specifically delay differential equations) that can be solved analytically, so that the density at a given time can be found as a function of the parameters μ, τ k_q, k_d and the initial density of replicative cells $R0$.

We have measurements of the total cell volume, i.e. $R(t) + Q_a(t) + Q_c(t)$ at several time points, for cell cultures with the following structure:

- 9 genetic designs, comprising 7 genetic interventions and two control designs.
- Between 1 and 4 clones implementing each design
- Two technical replicates for each clone.

Replicates of the same clone are biologically identical, though we expect some variation in measurements due to the experimental conditions. Clones with the same design are expected to be similar, with some degree of clonal variation that

may differ depending on the design. There is no prior information distinguishing the designs from each other, or distinguishing clones with the same design.

Given these assumptions a multi-level Bayesian statistical model is appropriate.

Measurement model

We used the following measurement model:

$$y \sim log\,normal(\log(\hat{y}(t, R0, \mu, \tau, k_q, k_d)), \sigma)$$

where

- $R0$, τ, k_q and k_d are vectors of clone-level parameters.
- μ is a model parameter representing the pre-treatment growth rate, which we assume is the same for all replicates.
- σ is an unknown log-scale error standard deviation.
- t is a vector of known measurement times (one per measurement).
- $\hat{y}$ is a function mapping parameter configurations to densities, under the delay differential equation assumptions laid out above.

We believe that the measurement error will be proportional to the true viable cell density for most measurements, motivating the use of the lognormal generalised linear model. However, we hypothesise that for cell densities below 0.3 this will not be the case, as for these measurements the error is dominated by factors that do not depend on the true cell density, such as impurities in the apparatus.

To allow the model to incorporate this postulated effect we use the following distributional model:

$$\sigma = exp(a_\sigma + b_\sigma * \min(0, \ln(\hat{y} - 0.3)))$$

In this equation the parameter b_σ represents the degree to which the log-scale measurement error increases or decreases as the true value gets lower than 0.3.

Code to generate figure 1:

```
import numpy as np
from matplotlib import pyplot as plt

plt.style.use("sparse.mplstyle")
y = np.linspace(0.04, 0.4, 20)
bs = [-0, -0.05, -0.1, -0.15, -0.2]
diff = np.array([np.log(yi/0.3) if yi < 0.3 else 0 for yi in y])
for b in bs:
    plt.plot(y, 0.2 + b * diff, label=str(b))
plt.legend(title="$b_\sigma$", frameon=False)
```

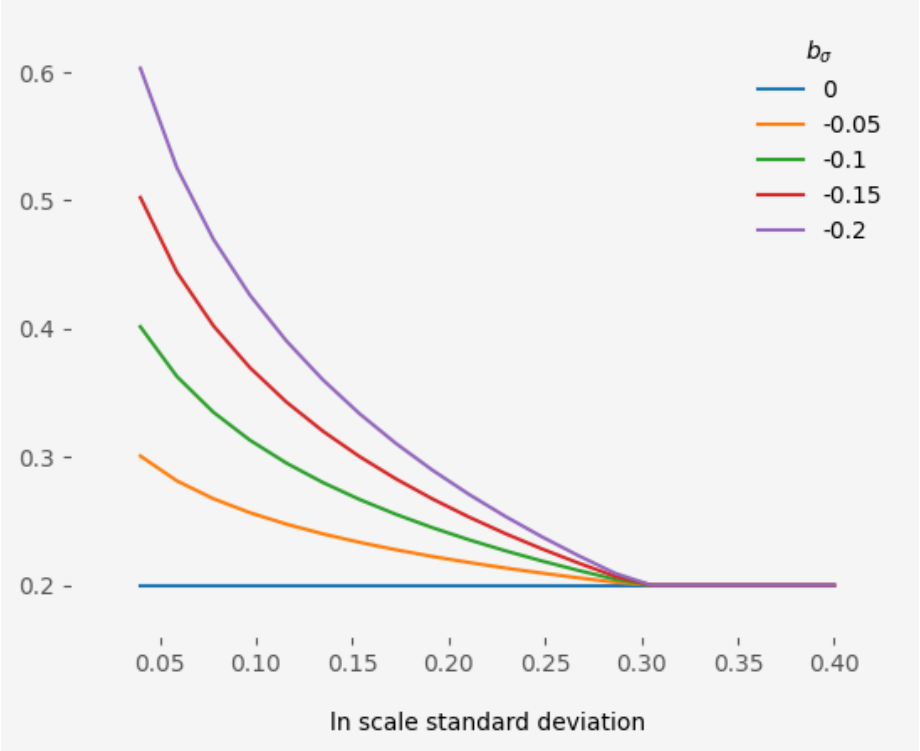

Figure 1: True density vs ln scale measurement standard deviation for a range of b_σ values

```
plt.xlabel("True density")
plt.xlabel("ln scale standard deviation")
plt.savefig("results/plots/y_vs_log_sd.png", bbox_inches="tight")
```

Design level parameters

The clone-level vectors τ_r and k_d are treated as determined by other parameters as follows:

$$\ln(\tau) = \tau const + d_\tau * X + c_{tau}$$
$$\ln(k_d) = dconst + d_d * X + c_d$$

In these equations

- $qconst$, $\tau const$ and $dconst$ are single unknown numbers representing the (log scale) mean parameter values with no interventions
- X is a matrix indicating which clones have which interventions
- d_τ and d_d are vectors of unknown intervention effects
- c_τ and c_d are vectors of unknown clone effects, representing random clonal variation.

The priors for the parameters d_τ and d_d were as follows:

$$d_\tau \sim N(0, 0.3)$$
$$d_d \sim N(0, 0.3)$$

To investigate whether the genetic interventions measurably affected the rate at which cells transition from the normal state R to the growth arrest state Q_a, we compared two different ways of modelling the clone-level vectors k_q. In the first model design M1, k_q is treated like τ and k_d, i.e

$$\ln k_q = qconst + d_q * X + c_q$$
$$d_q \sim N(0, 0.3)$$

In the second design M2, we assume that there are no design-level effects, i.e.

$$\ln k_q = qconst + c_q$$

Clonal variation parameters

The clone-level parameters c_τ, c_q and c_d have joint multivariate normal prior distribution:

$$\begin{aligned} [c_\tau, c_q, c_d] &\sim multi\,normal(\mathbf{0},\ _{\mathbf{c}} \cdot \Omega) \\ \Omega &\sim lkj(2) \\ _{\mathbf{c}}\&simlog\,normal(-2.1, 0.35) \end{aligned}$$

We used a multivariate normal distribution because we wanted to allow the possibility of correlations between clonal variation parameters: for example, if a certain clone has a very high death rate, this might predict a higher or lower rate of death-commitment.

The prior for σ_c is informative, and was chosen based on scientific knowledge so as to place 99% prior mass between 0.05 and 0.25.

In this equation lkj represents the Lewandowski, Kurowicka, and Joe distribution with shape parameter 2. See (**???**) for discussion of why this is an appropriate default prior for correlation matrices.

Informative priors for non-design parameters

Other unkowns have informative prior distributions based on scientific knowledge:

$$\begin{aligned} \mu &\sim log\,normal() \\ R0 &\sim log\,normal() \\ qconst &\sim normal() \\ \tau const &\sim normal() \\ dconst &\sim normal() \\ d_\tau &\sim normal() \\ d_d &\sim normal() \end{aligned}$$

Data representation

We believed that there should be no noticeable effects due to the "bok" intervention, as this intervention knocks out a gene that is very weakly expressed in the control case. To verify that this was the case we compared the results of

fitting the models M1 and M2 with and without distinguishing the "bok" design from the other designs.

Model comparison

To compare different models we calculated the approximate leave-one-timecourse-out log predictive density for models M1 and M2 on puromycin and sodium butyrate treatments using the python package arviz. See (Vehtari, Gelman, and Gabry 2017) for more about this model comparison method. For timecourses where the pareto-k diagnostic was higher than 0.7, indicating that the approximation was inaccurate, we refitted the model in order to find the exact leave-one-out log predictive density.

We further evaluated our models using graphical posterior predictive checks and by inspecting the modelled parameter values.

For both treatments, the two models' estimated predictive performance was very similar.

Results

Model comparison

We tested four model designs on two treatments, finding their estimated log predictive density, or elpd, using the semi-approximate leave-one-timecourse-out process described above.

The tables below show the results of this analysis. The models are listed in order of elpd. The column d_elpd

Sodium Butyrate:

model	elpd	se elpd	elpd	se elpd
m1_abc	-12.85248445	17.07466012	0	0
m2_abc	-14.43381192	17.12352247	1.581327475	2.709846992
m2_ab	-15.77080265	19.69527341	2.918318206	7.22210676
m1_ab	-17.94251902	18.15290027	5.090034574	2.689721148

Puromycin

model	elpd	se elpd	elpd	se elpd
m1_abc	-17.90297837	22.82799928	0	0
m2_ab	-19.66830572	23.43633536	1.765327355	5.316614552
m1_ab	-21.98055947	23.43963962	4.0775811	6.755786439

model	elpd	se elpd	elpd	se elpd
`m2_abc`	-30.56425074	35.08844529	12.66127237	19.49306278

We noted that the simplest and easiest to interpret model design `m2_ab` was easily within one standard error of the best leave-one-timecourse-out elpd for both treatments. We therefore chose to use this simpler model knowing that doing so did not entail a tangible sacrifice of predictive power.

Observed vs modelled timecourses

The figures in this section show observed timecourses for each design under the 15ug/mL puromycin treatment alongside 99% posterior predictive intervals, for each model and treatment.

The modelled and observed timecourses appear qualitatively similar in all cases, suggesting that none of the models are dramatically mis-specified.

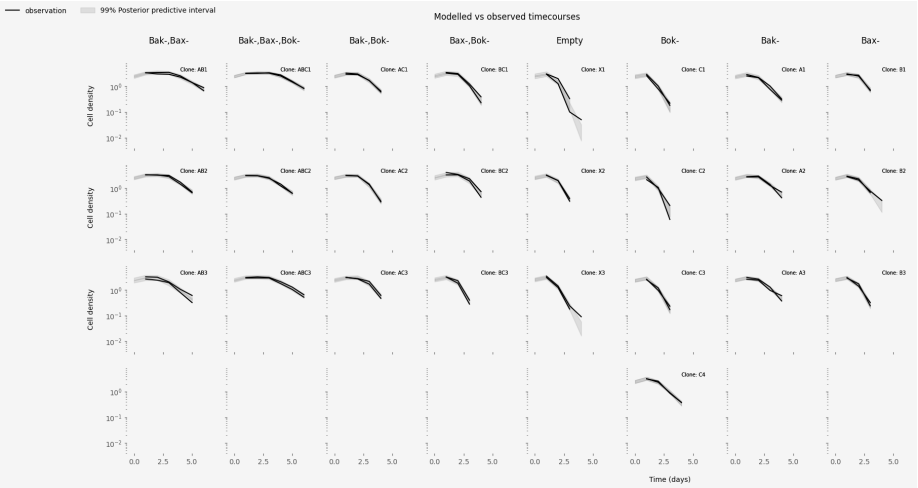

Figure 2: Puromycin treatment, design m1_abc

Posterior distributions of design parameters

The figures below plot the 2.5% to 97.5% marginal posterior intervals for the log-scale design-level effects, relative to the control experiment.

Some models are able to detect clear design effects with respect to the τ and k_d parameters. For example, for model `m2_ab` the posterior for the effect of design B on the delay parameter _d concentrates above zero for both treatments.

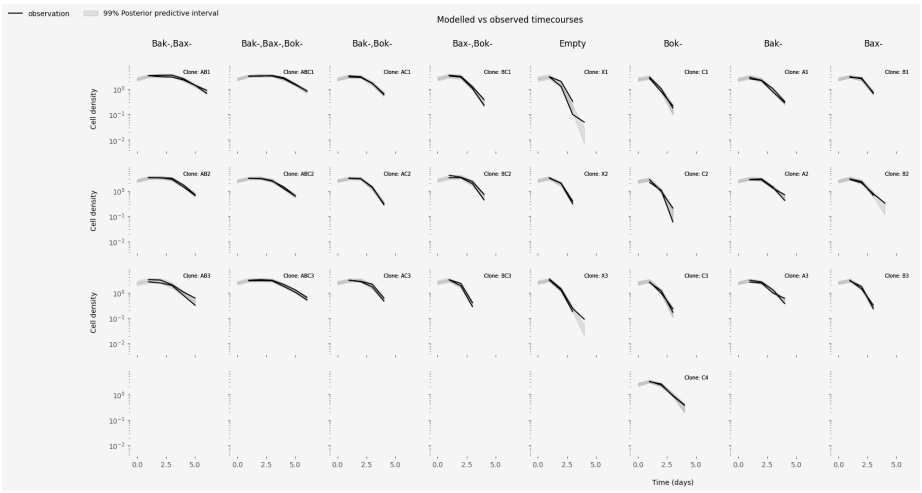

Figure 3: Puromycin treatment, design m1_ab

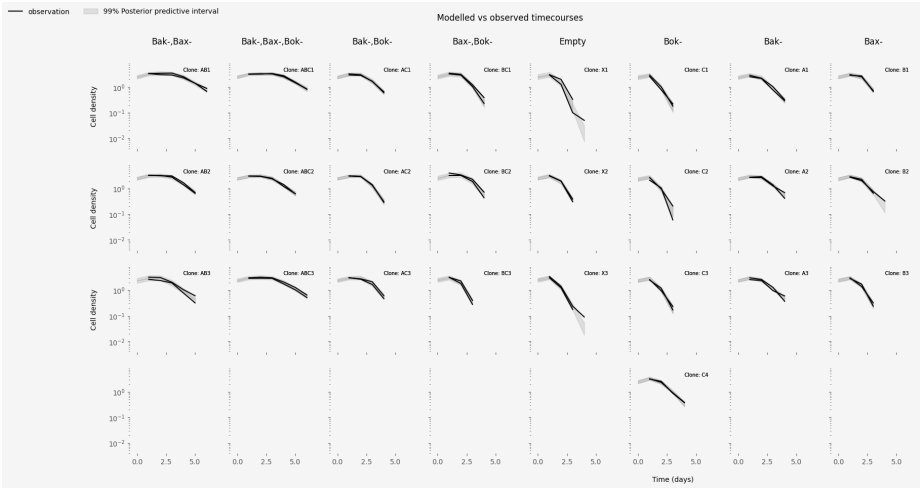

Figure 4: Puromycin treatment, design m2_abc

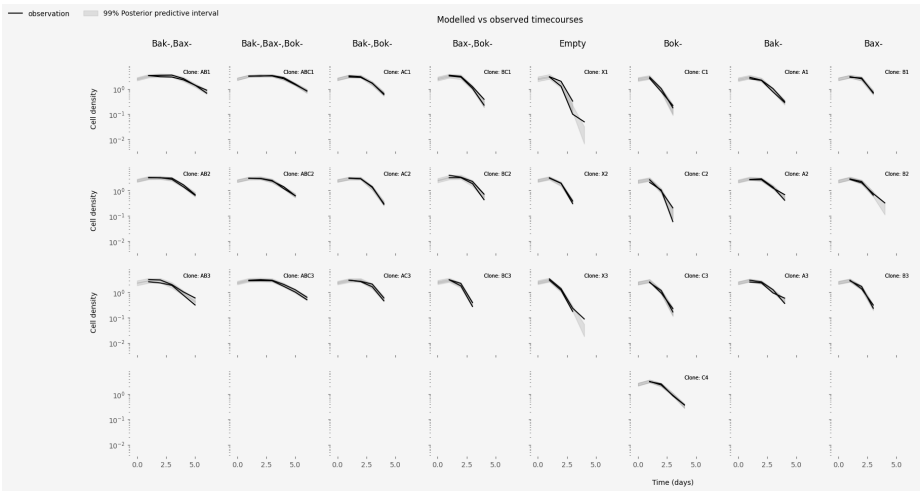

Figure 5: Puromycin treatment, design m2_ab

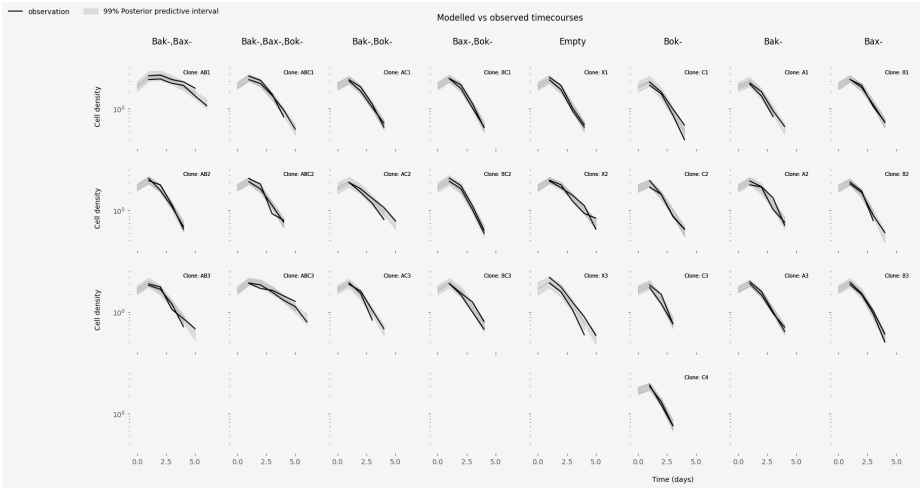

Figure 6: Sodium Butyrate treatment, design m1_abc

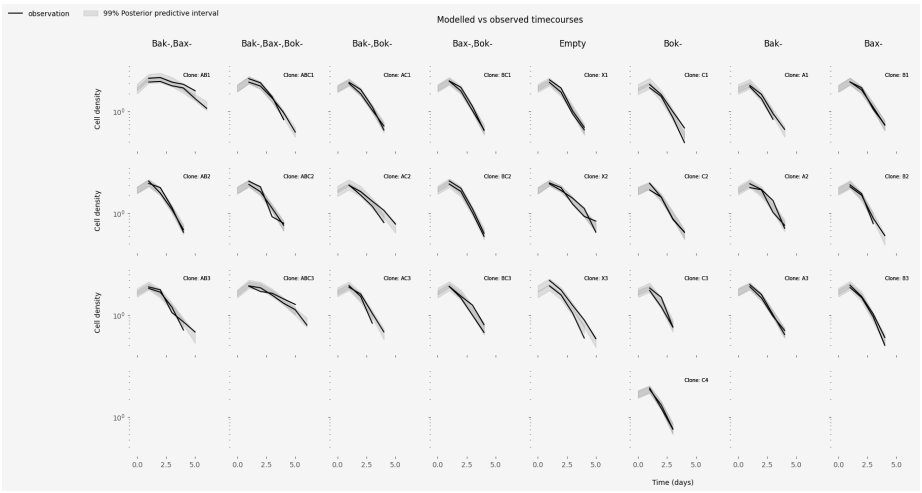

Figure 7: Sodium Butyrate treatment, design m1_ab

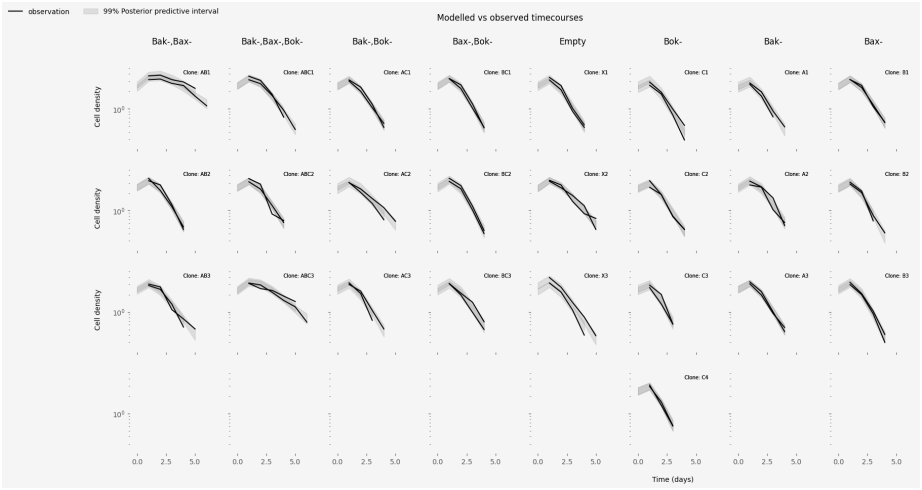

Figure 8: Sodium Butyrate treatment, design m2_abc

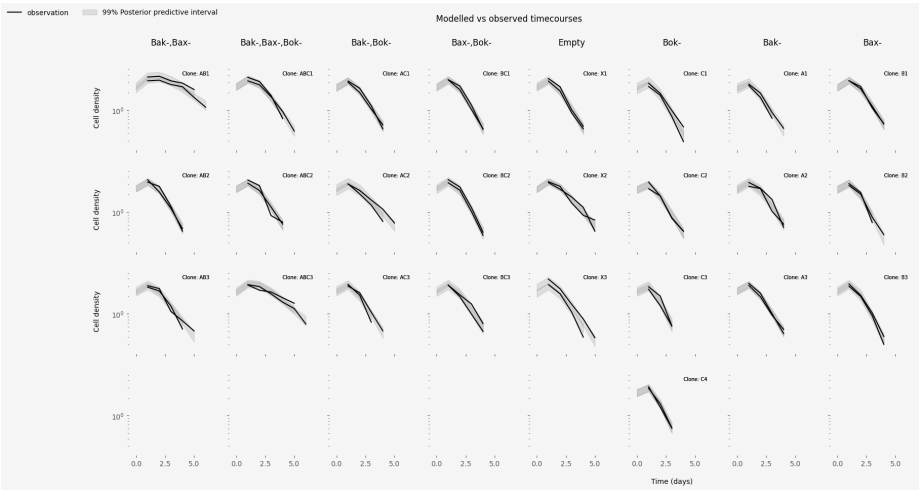

Figure 9: Sodium Butyrate treatment, design m2_ab

On the other hand, all of the posterior intervals for design-specific effects on the parameter k_q include zero, showing that none of the models can detect any impact from the designs on the speed of transition to growth arrest.

References

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC." *Statistics and Computing* 27 (5): 1413–32. https://doi.org/10.1007/s11222-016-9696-4.

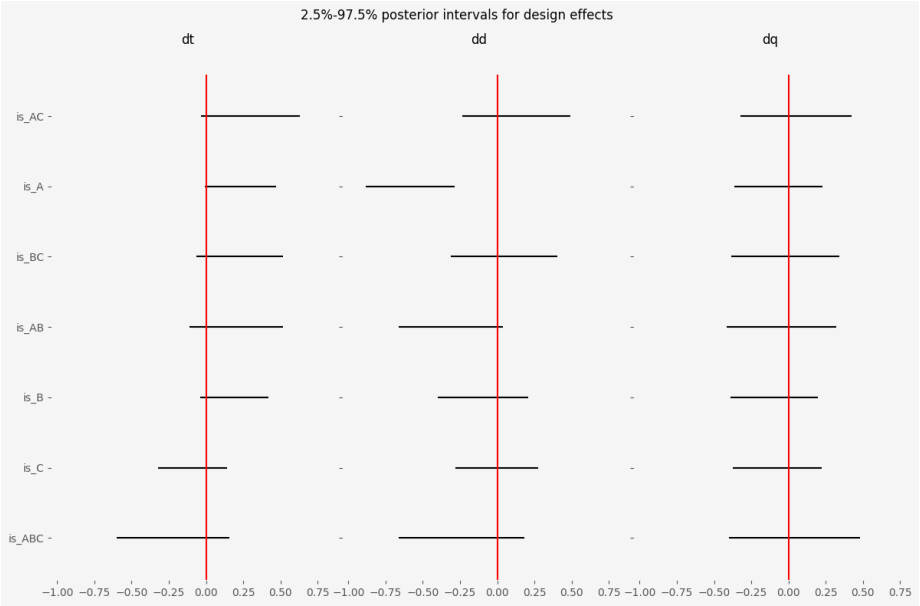

Figure 10: Posterior intervals for design level parameters, treatment 15ug/mL Puromycin, design m1_abc

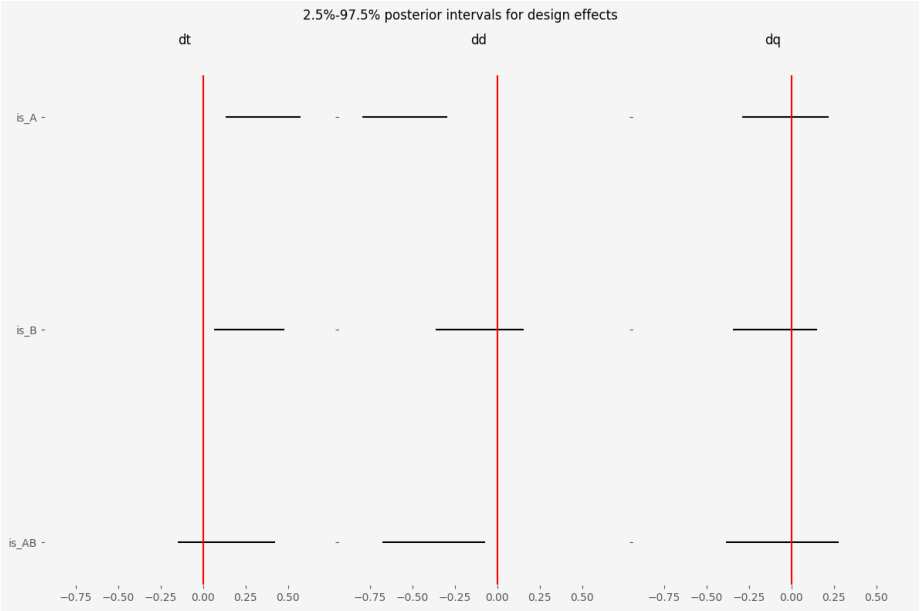

Figure 11: Posterior intervals for design level parameters, treatment 15ug/mL Puromycin, design m1_ab

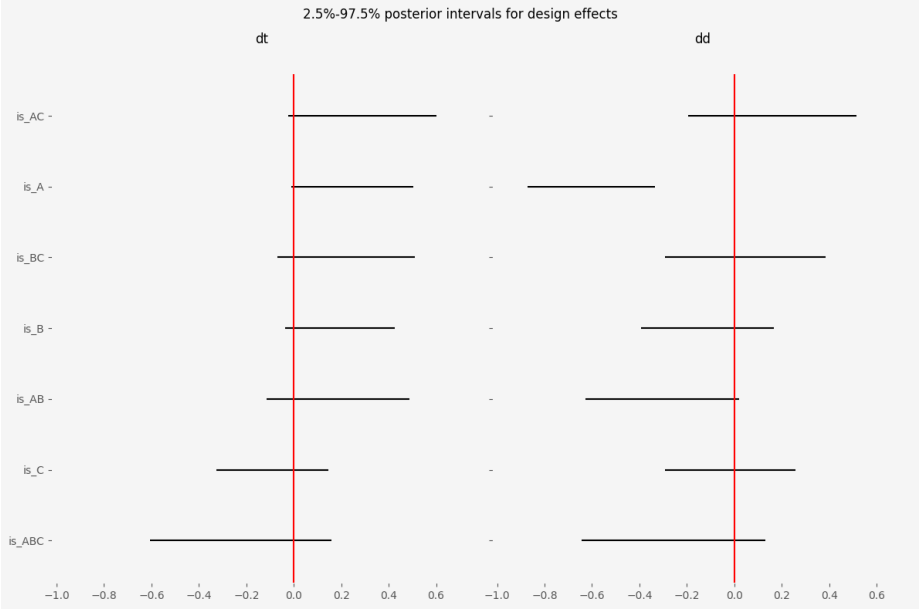

Figure 12: Posterior intervals for design level parameters, treatment 15ug/mL Puromycin, design m2_abc

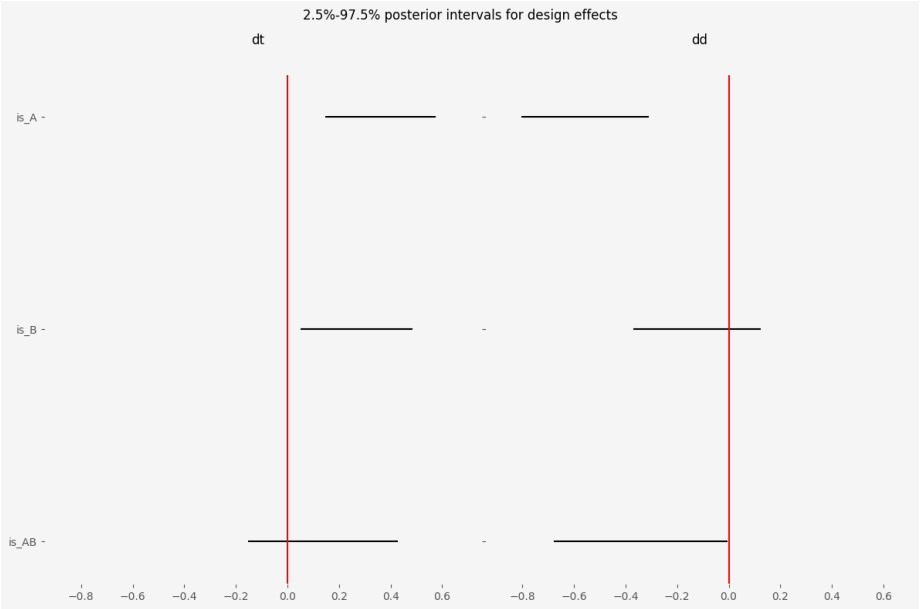

Figure 13: Posterior intervals for design level parameters, treatment 15ug/mL Puromycin, design m2_ab

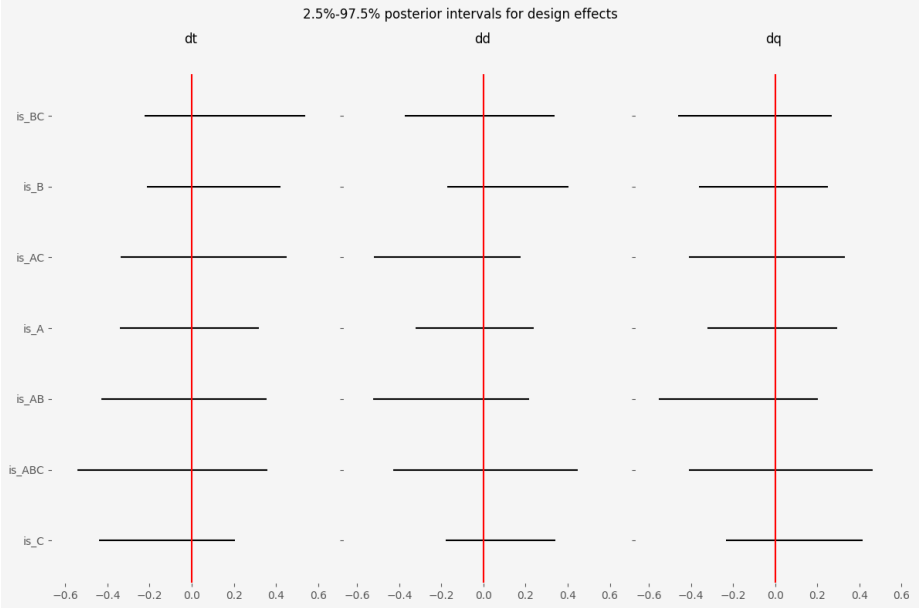

Figure 14: Posterior intervals for design level parameters, treatment 20mM Sodium Butyrate, design m1_abc

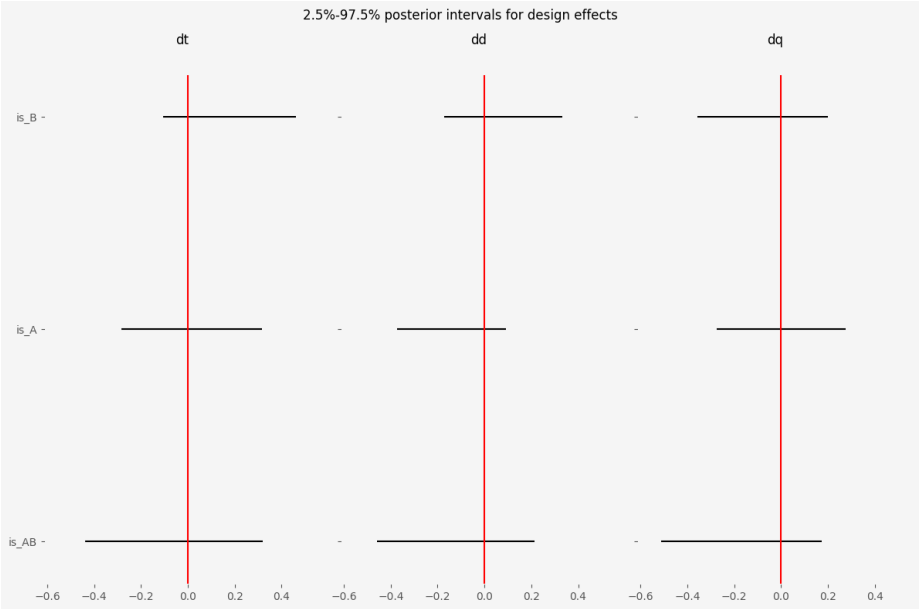

Figure 15: Posterior intervals for design level parameters, treatment 20mM Sodium Butyrate, design m1_ab

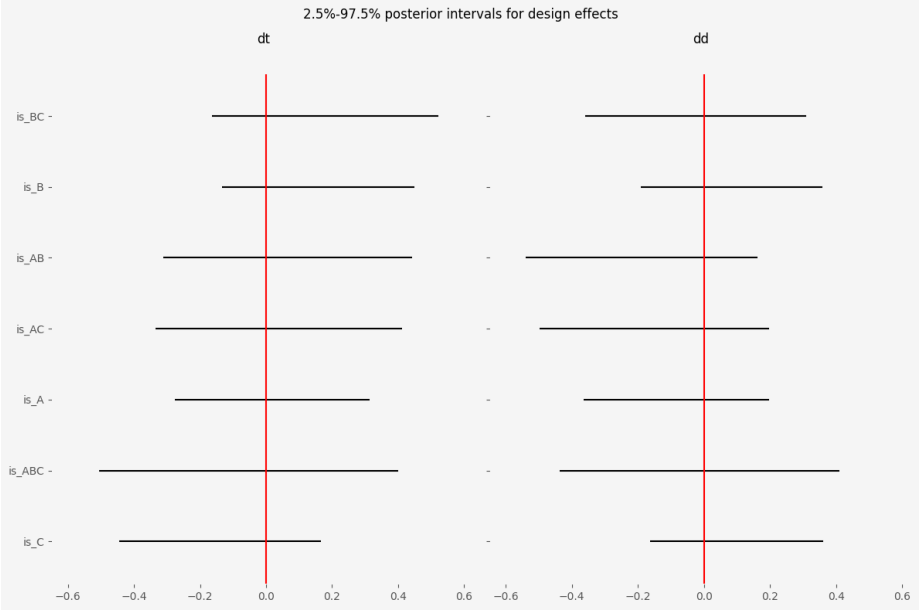

Figure 16: Posterior intervals for design level parameters, treatment 20mM Sodium Butyrate, design m2_abc

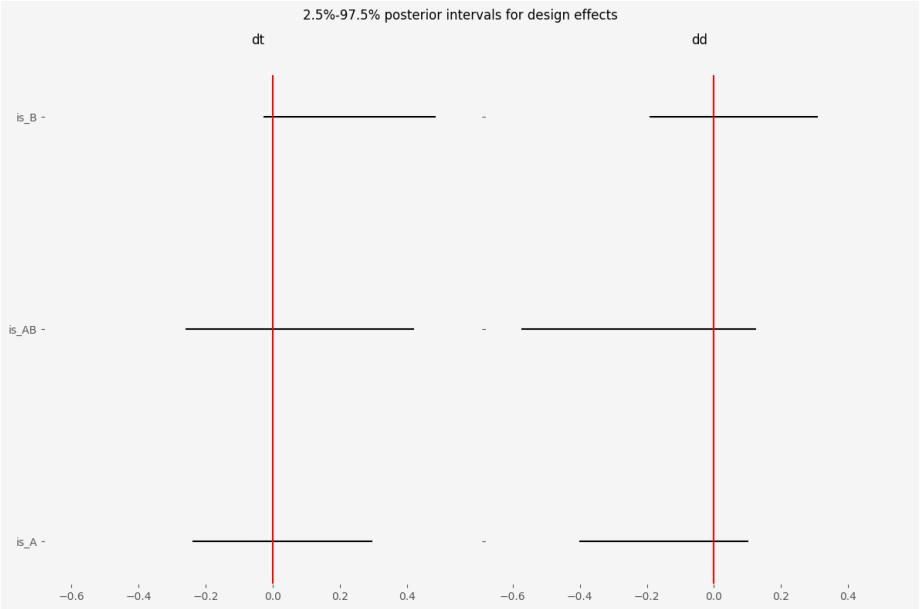

Figure 17: Posterior intervals for design level parameters, treatment 20mM Sodium Butyrate, design m2_ab